5.0 GUIDANCE TO REGULATORY AUTHORITIES, LABORATORIES AND PERMITTEES: GENERATING AND EVALUATING EFFECT CONCENTRATIONS

5.1 Steps for Minimizing Test Method Variability

This chapter provides the background and recommendations on WET test procedures related to sampling, conducting the toxicity test methods, and conducting the statistical methods. Implementing these recommendations should decrease or minimize WET test method variability, thereby increasing confidence to make regulatory decisions (see Figure 5-1). EPA stands behind the technical soundness of the current WET test methods. The critical steps in minimizing WET test method variability are (1) obtaining a representative effluent sample, (2) conducting the toxicity tests properly to generate the biological endpoints, and (3) conducting the appropriate statistical analysis to obtain powerful and technically defensible effect concentrations. Minimizing variability at each step increases the reliability of the WET test results. For example, factors that affect variability include sampling procedures; sample representativeness; deviations from standardized test conditions (e.g., temperature, test duration, feeding); test organisms; source of dilution water; and analyst experience and technique in conducting the toxicity tests properly (Burton et al. 1996).

5.2 Collecting Representative Effluent Samples

The goal of effluent sampling is to obtain a representative sample that reflects real-world biological responses. Factors affecting the representativeness of effluent samples may include the sampling location, frequency, and type (e.g., composite or grab), and sample volume, container, preservation methods, and holding time. Burton et al. (1996) concluded that the above factors considerably influence test result variability.

Effluent samples must be collected at a location that represents the entire regulated flow or discharge. Typically, the sampling site is designated in the discharge permit. As with sampling for any parameter, effluent samples should be collected from a location where the flow is turbulent and well-mixed. Additionally, effluent samples should be collected at a frequency that enables adequate characterization of the discharge over time (e.g., accounts for daily to seasonal changes and variations in effluent quality). Major facilities should conduct WET testing monthly or quarterly, while minor facilities should conduct WET testing semi-annually or annually.

Appropriate sample types should be collected to represent the effluent fully. When the effluent is variable, collecting composite samples may be necessary. When the effluent is less variable, grab samples may be sufficient (e.g., from long-term retention pond facilities).

Sample containers should be non-reactive so that they do not affect sample characteristics. Table II of 40 CFR Part 136 requires that toxicity test samples be collected in glass or plastic containers, as specified in the methods. Sufficient sample volume should be collected for the type of test being conducted, including the number of test dilutions. When samples are collected in Cubitainers®, headspace should be minimized.

Samples must be properly preserved. Part 136 of 40 CFR requires that samples for WET testing be cooled to 4°C when shipped off-site and between test sample renewals. Samples must be cooled during all phases of collection, transportation, and storage to minimize physicochemical changes. Samples must be tested within the specified maximum holding times before significant changes occur, such as volatilization or biological or chemical degradation. If samples are not tested within specified maximum holding times, the test is invalid and must be repeated by collecting a new effluent sample and conducting a new toxicity test to comply with the NPDES permit.

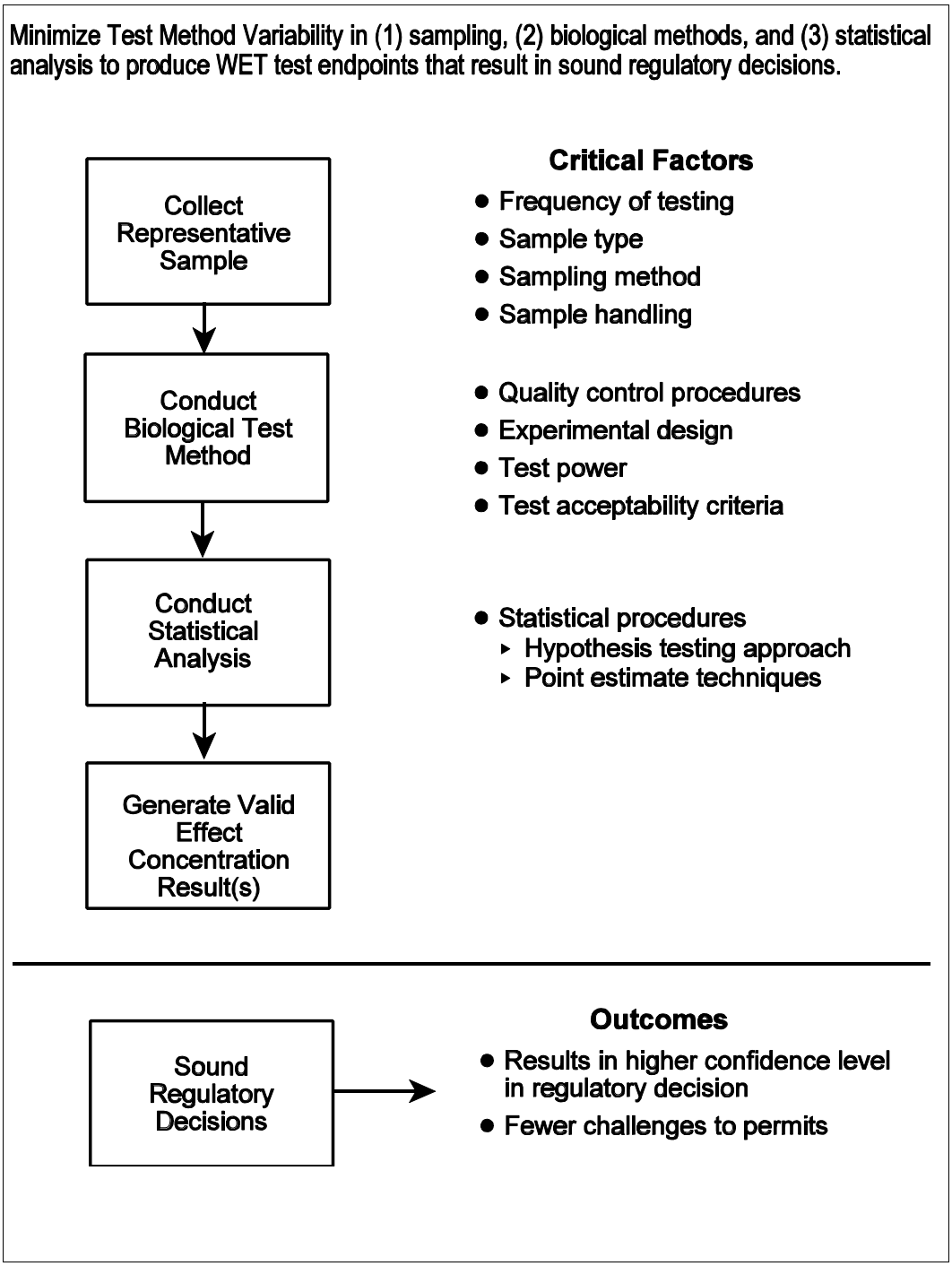

Figure 5-1. Steps to minimize WET test method variability.

5.3 Conducting the Biological Test Methods

Four main components of WET tests afford opportunities to control and minimize variability within tests and within and between laboratories: (1) quality control (QC) procedures; (2) experimental design; (3) test power; and (4) test acceptability criteria (TAC) beyond the minimum requirements specified in EPA's WET test methods.

5.3.1 Quality Control Procedures

Quality assurance (QA) practices for toxicity tests address all aspects of the tests that affect data quality. These practices include effluent sampling and handling, test organism source and condition, equipment condition, test conditions, instrument calibration, replication, use of reference toxicants, recordkeeping, and data evaluation. The EPA WET toxicity testing manuals specify the minimum requirements for each aspect. Regulatory authorities have the discretion to prepare and implement additional guidance beyond the minimum requirements specified in EPA's WET test methods.

An integral part of the QA program is quality control (QC). The QC procedures are the more focused and routine activities conducted under the overall QA program. An important QC component in WET testing is the requirement to conduct reference toxicant tests with effluent tests. The WET test methods outline when reference toxicant tests are to be conducted. (See sections on quality of test organisms in the manuals.) Reference toxicant testing serves two purposes: (1) determine the sensitivity of the test organisms over time; and (2) assess the comparability of within- and between-laboratory test results. Reference toxicant test results can be used to identify potential sources of variability, such as test organism health, differences among batches of organisms, changes in laboratory water or food quality, and performance by laboratory technicians. In the QA section of each promulgated test method (USEPA 1993, 1994a, 1994b), EPA recommends sodium chloride, potassium chloride, cadmium chloride, copper sulfate, copper chloride, sodium dodecyl sulfate, and potassium dichromate as suitable reference toxicants. The methods do not, however, specify a particular reference toxicant or the specific test concentrations for each test method.

The current characterization of WET test method variability is limited by the ability to quantify sources of within- and between-laboratory variability, because laboratories can use different reference toxicants and test concentrations for a particular method. Future evaluations of method variability would be greatly enhanced by having data to analyze from multiple laboratories for the same reference toxicant, the same dilution water at similar pH and hardness, and the same test concentrations. By standardizing reference toxicants, testing laboratories could compare test results, permittees and regulatory authorities could better compare and evaluate laboratories, and the data could be used to further quantify within- and between-laboratory test precision. Specification of the reference toxicant and test concentrations for a method across laboratories would provide a much larger and consistent data base to assess the comparability of within- and between-laboratory test results.

Standardizing reference toxicants and test concentrations has been discussed in the literature. For example, the chronic methods manual for West Coast species (USEPA 1995) specifies the reference toxicant and test concentrations for each test species. The Southern California Toxicity Assessment Group (SCTAG) is comprised of representatives from permittees, testing laboratories, regulatory authorities, and academic institutions that met to discuss technical aspects of WET testing (e.g., standardization of reference toxicants, control charts). The SCTAG (1996) prepared a report to standardize reference toxicants for the chronic freshwater test methods. This report evaluated an extensive data base of reference toxicant data. The report recommended specific reference toxicants and test concentrations for these methods. The SCTAG (1997) also prepared a QA/QC checklist to help toxicity testing laboratories establish and maintain appropriate data quality measures. Regulatory authorities should review these publications when standardizing reference toxicants.

The selection of reference toxicants and test concentrations should be based on specific criteria. The following criteria, recommended in the SCTAG report, provide an excellent basis for selecting standardized reference toxicants:

1. The toxicant should provide precise and reliable measures of toxicological sensitivity.
2. Toxicant disposal should not be legally or environmentally problematic.

3. The toxicant should produce a concentration-response effect for the test organism.
4. The toxicant should be quantifiable.
5. The toxicant should not pose an unacceptable health hazard to laboratory personnel.
6. The toxicant should be readily available.

Most recently, Warren-Hicks et al. (1999) recommended that national acceptance criteria be specified with upper and lower acceptance limits for reference toxicant test results, which all laboratories would need to achieve to obtain accreditation. Variability could decrease nationally if testing laboratories are provided with more detail on the evaluation and interpretation of reference toxicant control charts (APHA-AWWA-WEF 1998). For example, such guidance could describe how to evaluate test results within the warning limits. Both Environment Canada (1990, 2000) and APHA-AWWA-WEF (1998) have prepared guidance on evaluating control chart data. The Environment Canada (2000) report specifies using zinc as an inorganic reference toxicant and phenol as an organic reference toxicant for many aquatic tests. The report also specifies eight criteria for selecting specific reference toxicants.

1. Previous use
2. Availability in a pure form
3. Solubility
4. Stability in solution
5. Stability during storage
6. Ease of analysis
7. Stable toxicity with normal changes in qualities of laboratory water
8. Ability to detect abnormal organisms

Regulatory authorities may want to evaluate the above reports and the SCTAG reference toxicant recommendations for the chronic freshwater test methods. Regulatory authorities may also want to evaluate and recommend a standard reference toxicant and a specific concentration series for each acute and chronic test method using data from this guidance document.

5.3.1.1 Guidance Related to Quality Control Charts and Laboratory Audits

Ausley (1996) recommends some oversight of data quality, such as evaluating tests in meeting QC criteria, using randomization procedures, and operating in allowed reference toxicant ranges to ensure that QC procedures are properly implemented. Another integral component of QC is the maintenance of control charts for reference toxicants and effluents. Laboratories should provide regular review of control charts. EPA suggests keeping a control chart for each combination of test material, test species, test conditions, and endpoints with a maximum of 20 test results. Modern software makes accumulating data and reviewing key test statistics possible with relatively little effort. Elementary methods can identify problems contributing to variability. Laboratories should practice regular control charting of test PMSDs and control performance for all tests along with control charting of effect concentrations such as NOEC and point estimates for reference toxicants tests. Successive tests should be compared occasionally to detect repeated patterns, such as one replicate's being consistently higher or aberrant, or a trend over time. Time sequence plots of concentration means and standard deviations would be useful in this regard. Occasionally, a set of 5 to 20 tests, in which block positions (see Appendix A in USEPA 1994b) have been recorded, should be subjected to ANOVA for block or position effects. If such effects are significant or large, the laboratory should seek advice on randomizing the replicates and concentrations.

If a laboratory's CV exceeds the 75th percentile CV from Tables 3-2 through 3-4, EPA recommends calculating warning and control limits based on the 75th and 90th percentiles, respectively, of CVs for the method and endpoint (Tables 3-2 and 3-3 and Appendix Tables B-1 and B-2). For example, suppose the mean EC25 for a series of *Ceriodaphnia* chronic tests (Method 1002.0 with reproduction as the endpoint) conducted at one laboratory with reference toxicant is 1.34 g/L NaCl. Also suppose that the standard deviation of the EC25s for these tests is 0.85. The CV for this set of EC25s is thus 0.63. In Table 3-2, the 75th percentile of CVs for this test's reproduction endpoint is 0.45. Calculate the standard deviation corresponding to the 75th percentile CV, $S_{A.75} = 1.34 \times 0.45 = 0.60$. In Appendix Table B-1, the 90th percentile of CVs is 0.62 for this method and endpoint. Calculate $S_{A.90} = 1.34 \times 0.62 = 0.83$. Because the CV for this series of EC25s exceeds the 90th percentile reported in Table B-1, EPA recommends the following:

- Set control limits using $S_{A.90} = 0.83$,
- Set warning limits using $S_{A.75} = 0.60$,
- Promptly take actions to bring results within the control limits, and
- Attempt to bring results within the warning limits in 3-12 months.

If the CV for the set of EC25s is less than the 90th percentile reported in Table B-1, use that CV to set control limits. If the CV for the set of EC25s is less than the 75th percentile in Table 3-2, do not set warning limits using the latter value.

In addition, Burton et al. (1996) encourage regulatory programs to have a laboratory audit component to document the existence and effectiveness of a QA/QC program directed at toxicity testing, including analyst training and experience. Regulatory authorities should use the National Environment Laboratory Accreditation Program (NELAP) (USEPA 1999a) and routine performance audit inspections to evaluate individual laboratory performance. Inspections should evaluate the laboratory's performance with QC control charts based on reference toxicants, examine procedures for conducting the toxicity test procedures, and examine procedures for analyzing test results.

Regulatory authorities should develop a QC checklist to assist in evaluating and interpreting toxicity test results. Appendix E presents examples of State WET implementation procedures related to reviewing reference toxicant data and information on additional QA/QC criteria that have been developed and implemented. Regulatory authorities should also provide additional guidance related to the interpretation of QC control charts. This additional guidance could be that laboratories maintain control charts on within-test variability (e.g., PMSD) and use warning and control limits based on the 75th and 90th percentiles of CVs for the test method and endpoint.

5.3.2 Experimental Design

Experimental design includes randomizing the experimental units (i.e., treatments, organisms, replicates); establishing the statistical significance level (i.e., alpha level); and specifying the minimum numbers of replicates, test organisms, and treatments. Oris and Bailer (1993) recommend that test design and TAC be based, not only on a minimum level of control performance, but also on the ability to detect a particular level of effect (i.e., test power).

A Type I error (i.e., "false positive") results in the false conclusion that an effluent is toxic when it is not toxic. A Type II error (i.e., "false negative") results in the false conclusion that an effluent is not toxic when it actually is toxic. Power (1 - beta) is the probability of correctly detecting a true toxic effect (i.e., declaring an effluent toxic when it is in fact toxic). Acceptable values for alpha range from 0.01 to 0.10 (1 to 10 percent). The current EPA test methods recommend an alpha rate of 0.05 or 5 percent in the toxicity

testing manuals. Currently, EPA is preparing guidance on when an alpha rate of 0.01 or 1 percent would be considered acceptable (USEPA 2000a).

5.3.2.1 False Positives in WET Testing

The hypothesis test procedures prescribed in EPA's WET methods provide adequate protection against incorrectly concluding that an effluent is toxic when it is not. The expected *maximum* rate of such errors is the alpha level used in the hypothesis test. The hypothesis test procedure is designed to provide an error rate *no greater than* alpha when the default assumptions are met. The statistical flow chart provided with each EPA WET method identifies cases when default assumptions are not satisfied and, therefore, when data transformations or alternative statistical methods (e.g., a nonparametric test) should be used.

Alpha and beta are related (i.e., as alpha increases, beta decreases), assuming that the sample size (number of treatments, number of replicates), size of difference to be detected, and variance are held constant. The alpha and beta error rates depend on satisfying the assumptions of the hypothesis test. To ensure that statistical assumptions and methods are properly applied, testing laboratories should review the statistical procedures used to produce WET test results and other factors, such as biological and statistical quality assurance, and verify that test conditions and test acceptability criteria were achieved.

If a test is properly conducted and correctly interpreted, identifying any particular outcome as a "false positive" is impossible. An effluent that is deemed toxic may require that the permittee conduct additional toxicity tests to determine if toxicity is re-occurring. Even if no toxicity is demonstrated in follow-up tests, that does not rule out that the original toxic event was a true toxic spike in the effluent. False negatives, however, impact the environment by allowing the discharge of harmful toxicants without identification. This may occur because the toxic effects are not identified as statistically significant due to lack of test sensitivity (see Sections 5.3.3 and 6.4).

Measurement error should not affect the protection against false positives provided by hypothesis tests and confidence intervals when they are appropriately applied. Measurement error, in the case of WET test treatment mean values, likely consists largely of sampling errors (e.g., variability among organisms or containers), although errors in counting, weighing, and other procedures may also occur. Such sources of imprecision are implicitly accounted for in WET test statistical inferences, because the sample variance among the replicates within each treatment (dilution) is used for inference. The test "size" 1 - alpha will protect adequately against false positives. A larger variance among replicates, however, could make detecting real toxicity (i.e., false negatives) more difficult unless the number of replicates is increased to provide more test sensitivity and power, which will reduce the rate of false negatives.

5.3.2.2 False Negatives in WET Testing

For a given alpha, beta decreases (power increases) as the sample size increases and the variance decreases. Decreasing alpha from 0.05 to 0.01 without otherwise changing the hypothesis test will reduce the ability of the test to detect toxicity, that is, will reduce the power of the test. Thus, as alpha for the hypothesis test is decreased, there is an inevitable trade-off between the rate of false positives when toxicity is not present and the ability to detect toxicity when it is present (i.e., statistical power).

To limit within-test variability and thus increase power, EPA developed a minimum significant difference (MSD) criterion that must be achieved in the chronic West Coast marine test methods (USEPA 1995). The MSD is a measure of the within-test variability and represents differences between treatments and the control that can be detected statistically. Distributions of the MSD values of multiple tests for a specific reference toxicant and test method can be used to determine the level of test sensitivity achievable by a certain percentage of tests. Denton and Norberg-King (1996) analyzed several chronic test methods to quantify the effect size based on the existing toxicity test method experimental design and MSD distributions.

Denton and Norberg-King found when setting the beta error rate at 0.20 (power = 0.80), the effect size detected varies from at least a 15-percent reduction from the control response for the chronic red abalone larval development test to a 40-percent reduction from the control response for the chronic *Ceriodaphnia dubia* test. In this document, EPA has calculated power for each test method (see Section 5.3.3).

5.3.3 Test Power To Detect Toxic Effects

This section describes the statistical power and ability to detect toxic effects achieved by the current WET methods, as inferred from the WET variability data set used to develop this document. These inferences are approximate, because assumptions of normality and homogeneity of variance were not always satisfied.

Power can be characterized only by repeated testing. Power is an attribute not of a single test, but of a sequence of many tests conducted under similar conditions and with the same test design. Therefore, in this document, EPA used the sample averages for each laboratory's data set to characterize each laboratory. The following two parameters were required: (1) the mean endpoint response in the control (growth, reproduction, survival); and (2) the mean value of the error mean square (EMS) for tests.

EPA evaluated the ability to detect toxic effects using three approaches for each test method: (1) number of replicates required to detect a 25-percent difference from the control with power of 0.80; (2) percent difference from the control that can be detected with power of 0.80; and (3) power to detect a 25-percent difference from the control. All calculations are based on a one-sided, two-sample t-test at a level of 0.95 (alpha of 0.05). The power for a multiple comparison (Dunnett's or Steel's test) will be less than the power for this two-sample t-test.

Table 5-1 summarizes the results for this evaluation. Depending on the method, between 30 percent and 80 percent of the laboratories were able to detect a 25-percent effect for the sublethal endpoint consistently. Between 60 percent and 100 percent of the laboratories were able to detect a 33-percent effect.

To examine whether the upper bounds presented in Table 3-6 provide adequate test precision, EPA calculated an estimate of the power to detect a 25-percent effect on a sublethal endpoint when the PMSD equals the upper bound reported in Table 3-6. The upper bounds of the PMSD are shown in Table 3-6 in Chapter 3. At the lower PMSD bound, the power always exceeded 0.98. Tests with PMSD equaling the upper bound are not often able to detect a 25-percent effect. This finding does not mean that the upper bound is ineffective. The PMSD varies between tests, and each laboratory has a distribution of PMSDs. To avoid exceeding this upper bound often, a laboratory would have to achieve substantially lower PMSDs in most tests.

5.3.3.1 Attainment of the PMSD Related to Power

The power of the current experimental design could be reevaluated by comparing it to alternative designs that use increased number of replicates or number of test concentrations (Chapman et al. 1996). In this document, EPA found that about half of the laboratories in the data set were able routinely to detect a 25-percent difference between control and treatment. About two-thirds of the laboratories could routinely detect a 33-percent difference (Table 5-2). For example, mere attainment of the 90th percentile PMSD values shown in Table 3-6 will not ensure the ability to detect a 25-percent effect (Table 5-2). If every acceptable test has a PMSD below that upper bound, however, the average PMSD will be lowered. Based on the within-laboratory variability of PMSD,[1] the average PMSD likely will be substantially lower than the upper bound in Table 3-6, if *most* tests conducted by a laboratory are to have acceptable PMSDs.

[1] The average CV for PMSD is one-third to one-half its mean in commonly used methods.

Table 5-1. Tests for Chronic Toxicity: Power and Ability To Detect a Toxic Effect on the Sublethal Endpoint

Test Method	No. Labs	No. Labs with Power		Power (Range)	No. Labs Having Power at Least 0.8 To Detect Effect of		Effect Detected with Power 0.8 as Percent of Control Mean (Range)
		0.8	0.5		25%	33%	
1000.0 Fathead Minnow	19	6	14	0.21 - 1.00	6	13	8.2 - 62
1002.0 *Ceriodaphnia*	33	10	29	0.38 - 1.00	10	19	14 - 45
1003.0 Green Alga	9	7	8	0.33 - 0.99	7	8	13 - 49
1004.0 Sheepshead Minnow	5	4	5	0.77 - 1.00	4	5	8.6 - 26
1006.0 Inland Silverside	16	7	13	0.23 - 0.97	7	12	17 - 59
1007.0 Mysid (growth)	10	5	8	0.21 - 0.91	5	8	21 - 70

Note: Power was calculated for a two-sample, one-sided t-test at alpha = 0.05, for a 25-percent difference from the control. Effect size detected was calculated for the same test using power 0.80. Calculations used the average EMS from all tests at each laboratory and the minimum number of replicates reported for those tests. Calculations assumed that the parametric mean and variance equal the corresponding sample estimates. They also assumed approximate normality of means and homogeneity of variance. Because these assumptions may be violated, the results here are approximate. By saying "detect a 25-percent difference from control," this alternative hypothesis is intended: (control mean - treatment mean) > 0.25 × control mean.

Table 5-2. Power To Detect a 25-Percent Difference from the Control at the 90th Percentile PMSD

Chronic Method	Replicates	90th Percentiles of PMSD	Three Treatments		Four Treatments		Five Treatments	
			alpha = 0.05	alpha = 0.05/3	alpha = 0.05	alpha = 0.05/4	alpha = 0.05	alpha = 0.05/5
1000.0 Fathead Minnow	3	35	0.39	0.25	0.39	0.19	0.39	0.15
	4	35	0.41	0.30	0.42	0.26	0.43	0.23
1002.0 *Ceriodaphnia*	10	37	0.39	0.31	0.41	0.30	0.43	0.30
1003.0 Green Alga	3	35	0.39	0.25	0.39	0.19	0.39	0.15
	4	35	0.41	0.30	0.42	0.26	0.43	0.23
1004.0 Sheepshead Minnow	3	23	0.72	0.69	0.72	0.62	0.73	0.55
	4	23	0.73	0.71	0.74	0.68	0.75	0.66
1006.0 Inland Silverside	3	23	0.72	0.69	0.72	0.62	0.73	0.55
	4	23	0.73	0.71	0.74	0.68	0.75	0.66
1007.0 Mysid	8	32	0.48	0.41	0.50	0.40	0.52	0.40

Notes: Values are rounded to two significant figures. Number of treatments is the number of concentrations compared with the control in the hypothesis test. The calculations assumed (1) the usual assumptions of the test are satisfied (approximate normality, homogeneity of variances); and (2) equal replication in treatments and control. Because these assumptions may be violated, the results here are approximate. Because the MSD/mean implies a value for [root (error mean square)/mean], the latter could be calculated from the MSD, Dunnett's critical value, and the number of replicates, and then used in a calculation of power. Calculations apply to a one-sided, two-sample t-test of equal means, assuming equal variances and equal replication, with hypotheses H_o:{control mean - treatment mean = 0} versus H_a:{control mean - treatment mean > 0.25 × control mean}. The power achieved by Dunnett's multiple comparison procedure will lie between the two-sample power at alpha = 0.05 and that for alpha = 0.05/(no. of treatments).

Testing laboratories and permittees can examine the EMS or MSD in Tables B-14 and B-15 (Appendix B) to estimate the ability of a WET test to detect toxic effects. Some regulatory authorities may require a comparison between the control and the receiving water concentration, which requires a two-sample, one-sided t-test. Others may require the multiple comparisons procedure described in the EPA WET methods (Dunnett's or Steel's tests, one-sided, with alpha of 0.05). The power of Dunnett's procedure falls between the power of the one-sided, two-sample t-test with alpha of 0.05 and alpha of 0.01, assuming that no more than five toxicant concentrations are compared to a control. The power of Steel's procedure will be related to, and should usually increase with, the power of Dunnett's procedure and the t-tests. Tables B-14 and B-15 in Appendix B also provide an appropriate guide to achieving power using a nonparametric test.

Recently, the State of Washington (1997) issued guidance specifying an acute and chronic statistical power standard to be achieved for compliance testing. EPA's sediment toxicity testing manuals (USEPA 1994c, USEPA 2000) include power curves for various numbers of experimental units, CV ranges, and associated alpha and beta levels. Sheppard (1999) is a good source to provide a simple explanation of how power helps determine how large a sample should be. Additional information on power may be obtained at: http://www.psychologie.uni-trier.de:8000/projects/gpower/literature.html.

EPA recommends that regulatory authorities specify in their State WET implementation procedures that individual test results achieve a level of within-test sensitivity by using the upper and lower PMSD test sensitivity bounds (see Section 6.4). To achieve the test sensitivity bounds, testing laboratories may need to minimize within-test variability (e.g., EMS) or increase the number of replicates tested, or both. If laboratories cannot achieve PMSD values of less than 25 percent for the toxicity test methods that require a minimum of only three replicates (Methods 1000.0, 1004.0, 1006.0), then the numbers of replicates may need to be increased. Appendix B (Section B.4) provides information related to the number of replicates needed and discusses the relationship between test power and effect size achieved. The magnitude of the effect size achieved relates to the test sensitivity.

5.4 Test Acceptability Criteria

EPA test methods have specific TAC that the effluent and reference toxicant tests must meet. A test is considered invalid if the TACs are not met. The recommended test conditions for each test method specify the minimum requirements and the TAC. For example, control survival must be 80 percent or greater and average control reproduction at least 15 young per surviving female in the chronic *Ceriodaphnia dubia* survival and reproduction test.

The chronic West Coast marine methods (USEPA 1995) require additional TAC. For example, to limit the degree of within-test variability, the methods specify a maximum allowable value for PMSD (see Section 5.3.2 on experimental design). Some States have additional TAC in their State WET implementation policies. North Carolina (1998) for example, requires that the chronic *Ceriodaphnia dubia* analyses meet an additional TAC of complete third brood neonate production by at least 80 percent of the control organisms and that the control reproduction CV be less than 40 percent.

Additional TAC might be specified to minimize variability among replicates. Variability of any toxicity test result is influenced by the number of replicates used, number of organisms tested, and variability among replicates at each test concentration and the control. Variability among replicates has been quantified by treatment CV, EMS, or MSD. The application of a maximum acceptable value for CV or MSD helps ensure adequate laboratory QA/QC and increases the reliability of submitted data. One benefit of requiring a maximum allowable within-test variability limit is that laboratories will improve culturing, test handling, and housekeeping, which are usually incorporated into the laboratories' standard operating procedures. For example, the CV requirement might be incorporated directly into the NPDES permit. Sample EPA Region 6 permit language reads:

1. *The coefficient of variation between replicates shall be less than or equal to 40 percent in the control.*

2. *The coefficient of variation between replicates shall be less than or equal to 40 percent at the instream waste concentration (IWC).*

3. *Test failure may not be construed or reported as invalid due to a CV of greater than 40 percent. A repeat test shall be conducted within the required reporting period if any test is determined to be invalid.*

Occasionally, statistical analyses indicate a test failure when as little as 15-percent mortality has occurred in a test dilution. Permit language has been developed to address this occurrence, as in the following example:

> *If all TAC conditions are met and the percent survival of the test organism is greater than or equal to 80 percent (in a chronic test) or 90 percent (in an acute test) in the critical dilution concentration and all lower dilution concentrations, the test shall be considered to be a valid test, and the PERMITTEES shall report an NOEC of not less than the critical dilution for the discharge monitoring report (DMR) reporting requirements.*

Regulatory authorities may consider providing guidance or imposing additional TAC, such as those implemented by EPA Region 6 or like some States have implemented (North Carolina 1998, Washington 1997). Appendix E provides additional examples of States that have implemented further guidance on WET QA/QC procedures and TAC. Warren-Hicks (1999) also recommended that additional national TAC be established for each test method (e.g., upper and lower bounds on the MSD). Therefore, EPA recommends that regulatory authorities require that additional TACs be implemented in permits to minimize within-test variability and increase test sensitivity (see Section 6.4 and Appendix C for sample permit language).

5.5 Conducting the Statistical Analysis To Determine the Effect Concentration

EPA test methods currently recommend two statistical approaches to estimate a chemical or effluent concentration for each biological effect endpoint (e.g., survival, growth, and reproduction). One approach is to derive the NOEC by hypothesis testing, which equates biological significance with statistical significance. The second approach is to estimate an effect concentration that reduces the control response by 25 percent for chronic methods. The expanded use of WET tests in the NPDES program has brought increased attention to the statistical analysis of toxicity test data. A common goal for both regulatory authorities and permittees is to confirm that the effect concentrations were derived correctly using the appropriate analysis approaches. Reliable effect concentrations lead to increased confidence in the data used for making regulatory decisions, such as determining reasonable potential, deriving a permit limit or monitoring trigger, and generating self-monitoring test results.

Another important consideration in conducting statistical analyses is the inconsistent use of statistical programs. The proliferation of statistical packages has been helpful in data analysis; however, these packages also can result in the misapplication of the statistical methods. APCA-AWWA-WEF (1998) cautions the user to confirm the results of each analysis with each package before accepting them. The data user is responsible for evaluating all data submitted to the regulatory authorities.

The 1995 SETAC Pellston Workshop discussed unresolved scientific issues and highlighted significant research needs associated with WET testing. The attendees recommended the following:

> *Immediately instigate studies to evaluate improvements in the statistical analysis of WET test data. These studies should include, but not necessarily be limited to, the following activities:*

(a) investigate the implications of concurrent application of NOEC/MSD, tests of bioequivalence, and ECp estimators (Chapman et al. 1996a).

In response to this recommendation, EPA began projects to evaluate the bioequivalence approach and additional point estimate models for the WET program. At present, two test methods are being used for this evaluation: (1) the chronic *Ceriodaphnia dubia* survival and reproduction tests and (2) the giant kelp germination and germ-tube length test with reference toxicants.

The bioequivalence approach poses the following question: Do the mean responses of the effluent concentration and the control differ by more than some amount? For example, the control response and the response at the critical effluent concentration (i.e., instream waste concentration) must differ by no more than a fixed value in order to accept the hypothesis of no significant difference (i.e., no toxicity). This approach could address the concern that an imprecise test might not detect toxicity when toxicity is present or that a small but statistically significant effect would detect toxicity that may not be biologically important. Some researchers have suggested that the bioequivalence approach could provide a positive incentive for dischargers to produce test results with lower within-test variability to demonstrate that no toxicity occurs at a level greater than a biologically (bioequivalence approach) significant amount (Shukla et al. 2000, Wang et al. 2000).

Bailer et al. (2000) evaluated the proposed regression-based estimators with the current EPA point estimate models. They found that it appears reasonable to incorporate parametric estimation models in the WET program. Bailer et al. (2000) concluded that these models are appropriate for all response scales (i.e., dichotomous, count, and continuous) and can incorporate monotonicity without bias. However, confidence intervals still need to be developed for these parametric models.

In this document, EPA has not recommended either the bioequivalence or additional point estimate models to supplement the current statistical approaches as described in the testing manuals. An independent, peer-reviewed workshop should be convened to evaluate the benefits of these alternative statistical approaches to enhance the statistical approaches currently applied.

5.6 Chapter Conclusions

In this chapter, EPA provides guidance to permittees and testing laboratories on collecting representative effluent samples, conducting the biological test methods, and evaluating the statistical analyses. EPA recommends that States implement the lower and upper PMSD test sensitivity bounds to achieve an acceptable level of test sensitivity and minimize within-test variability (see Section 6.4). EPA also provides guidance to permittees and testing laboratories on the number of replicates required to achieve the PMSD bounds. Testing laboratories should maintain and evaluate both effluent and reference toxicant data using a measure of within-test variability such as the PMSD.

Permittees and toxicity testing laboratories may need to increase replication in order to reduce PMSD below the upper bound. Table B-15 can be used for initial planning of replication, given knowledge of typical values of the error mean square (EMS) or MSD and the number of concentrations used in the multiple comparison hypothesis test. To ensure that all PMSD values fall below the upper bound in Table 3-6, a laboratory would select the largest EMS value experienced in its past testing.

EPA recommends that testing laboratories require a minimum of four replicates for the fathead minnow, sheepshead minnow, and inland silverside chronic test methods (Methods 1000.0, 1004.0, and 1006.0, respectively). Four replicates are needed to execute the statistical flow chart when a nonparametric test is needed. Three replicates are also sometimes insufficient to keep PMSD below the recommended upper bound. In addition, four replicates are needed to help achieve the upper PMSD bound.

This page intentionally left blank.